(Un)likelihood Learning for Interpretable Embedding

Jiaxin Wu[1], Zhijian Hou[1], Zhixin Ma[2] and Chong-Wah Ngo[2]
[1]City University of Hong Kong
[2]Singapore Management University

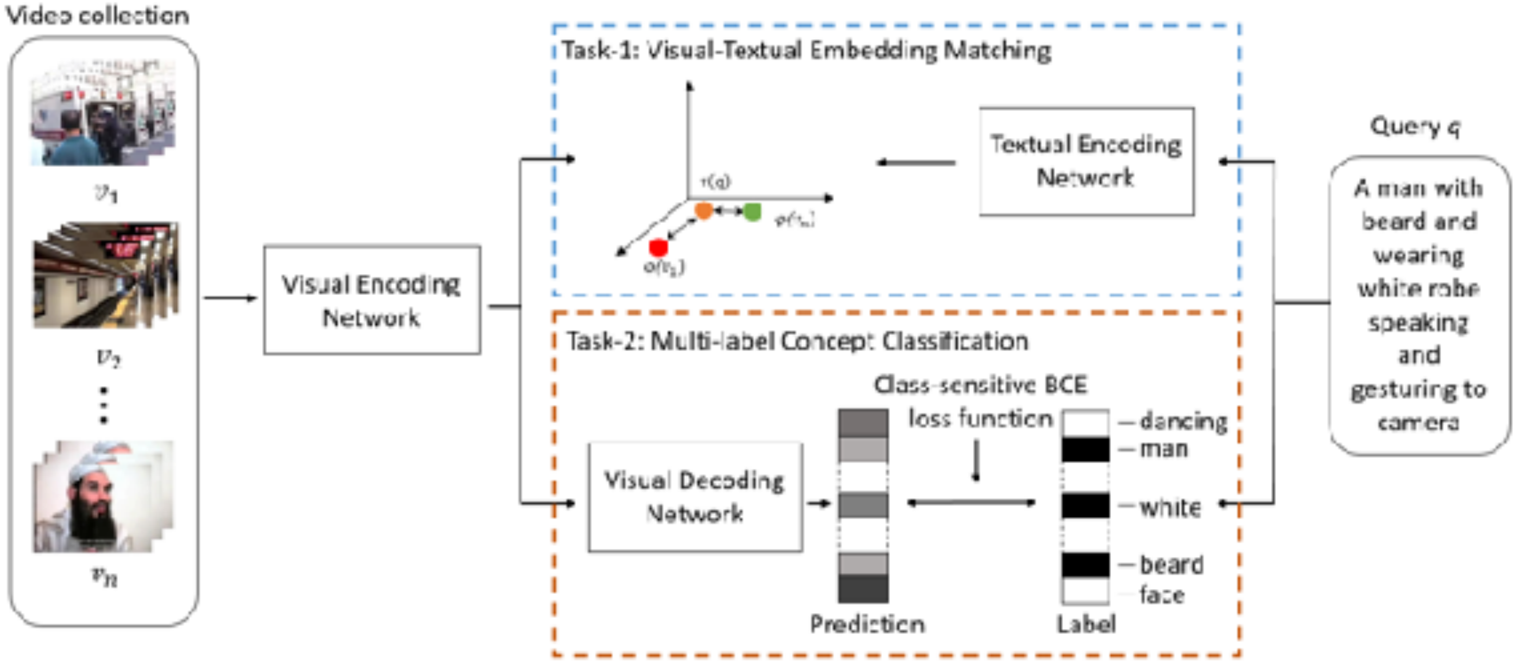
TRECVID 2021 Workshop

Outline

- Interpretable embedding and overlooked issue
- Unlikelihood learning for interpretable embedding
- Submitted runs and analysis
- Summary

Interpretable embedding (Dual-task model)

- Main idea: equip embedding search with interpretability.

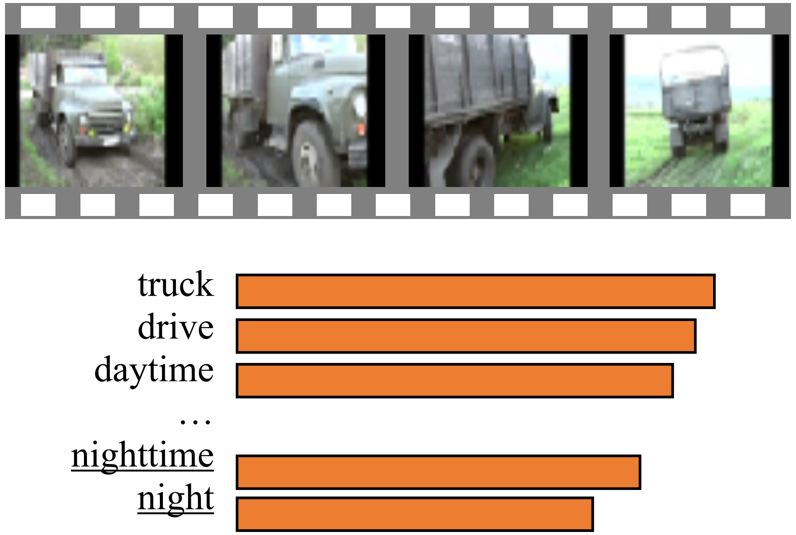


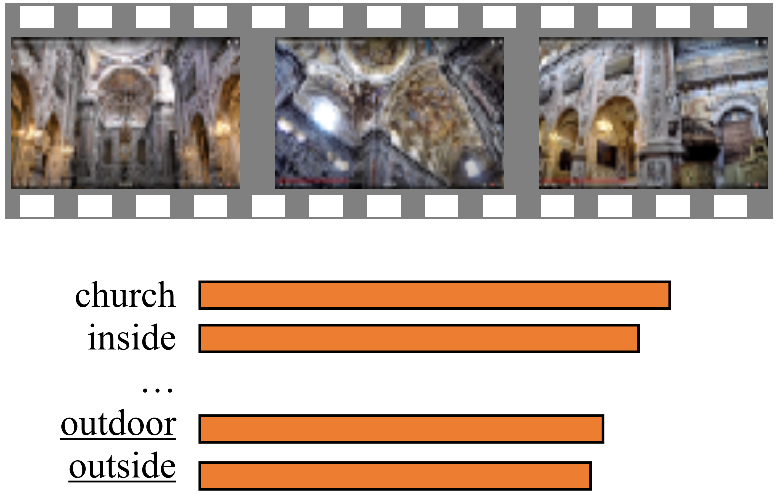
Wu and Ngo, Interpretable Embedding for Ad-hoc Video Search, *ACMMM*, 2020

Overlooked Issue: Inconsistent Interpretation

- Contrary concepts are simultaneously decoded for visual embeddings
- Hurt representation learning and retrieval performances

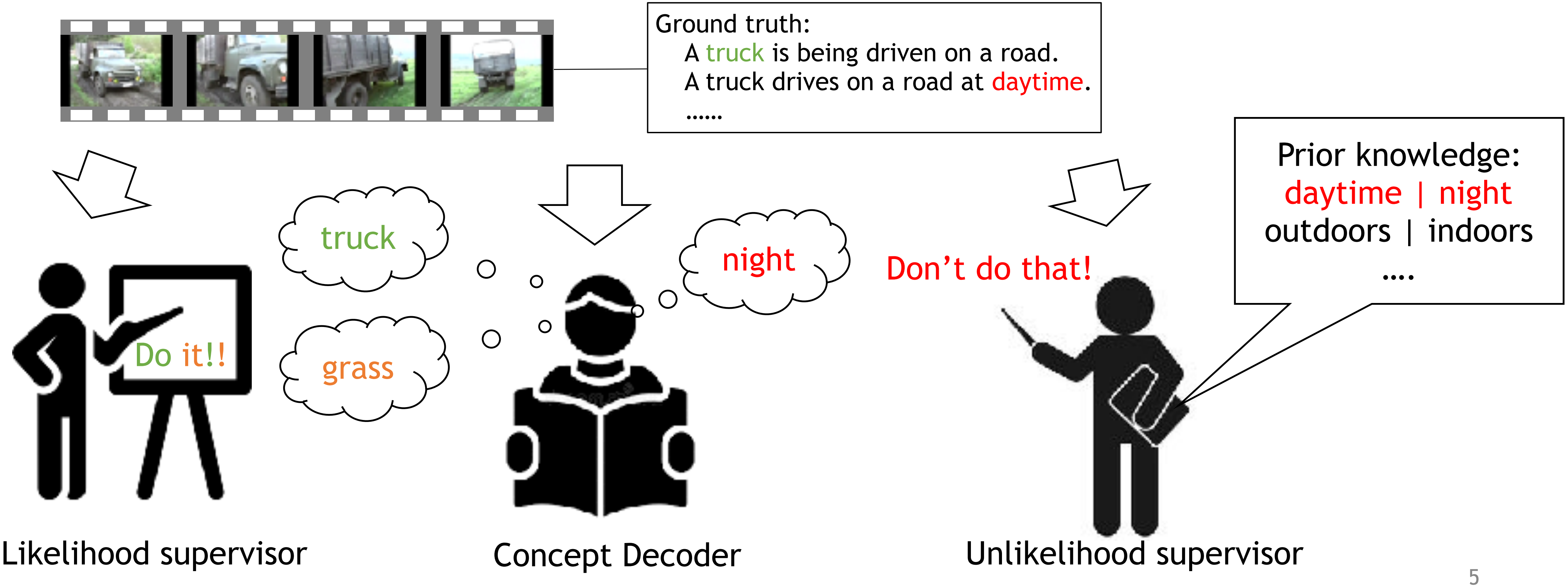
Visual embedding interpretation





How to generate consistent interpretation

- Two “supervisors” (Likelihood and Unlikelihood)



Likelihood learning

- Goal: recover the concepts in the annotated label.
- Obstacles: sparse and incomplete label.
- Propose class-sensitive BCE loss.

$$
Loss_{BCE}(\hat{p}, p) = \lambda \frac{1}{\sum_i^n p_i} \sum_i^n p_i BCE(\hat{p}_i, p_i)
+ (1-\lambda) \frac{1}{\sum_i^n (1-p_i)} \sum_i^n (1-p_i) BCE(\hat{p}_i, p_i),
$$

$$
BCE(\hat{p}_i, p_i) = -[p_i log(\hat{p}_i) + (1-p_i) log(1-\hat{p}_i)].
$$

Ground truth:
A truck is being driven on a road.
A truck is moving on a road.
A truck drives on a road at daytime.
……

Sparse and incomplete ground truth

	truck	cat	drive	grass	talk	…	water	daytime	road	
p	1	0	1	0	0	…	0	1	1	$\in \{1,0\}^{10,000+}$

Concept prediction

	truck	cat	drive	grass	talk	…	water	daytime	road	
$\hat{p}$	0.99	0.15	0.97	0.75	0.32	…	0.12	0.89	0.96	$\in \mathbb{R}^{10,000+}$

Unlikelihood Learning (UL)

- Goal: suppress the probabilities of contradicting/exclusive concepts.
- Prior knowledge: WordNet antonym[1].
- Obstacles: Context, globally/locally exclusive
- Propose new UL loss function inspired by [2,3].

$$Loss_{UL}(\hat{p},p) = \frac{1}{\sum_i^n p_i}\sum_i^n -p_i \sum_{t\in T_i} log(1-\hat{p}_t)*(1-p_t)$$

Ground truth

index	1	2		i		j	$j+1$		t		n
p	0	0	…	1	…	1	0	…	1	…	0
	truck	cat		man		indoors	outdoors		woman		road

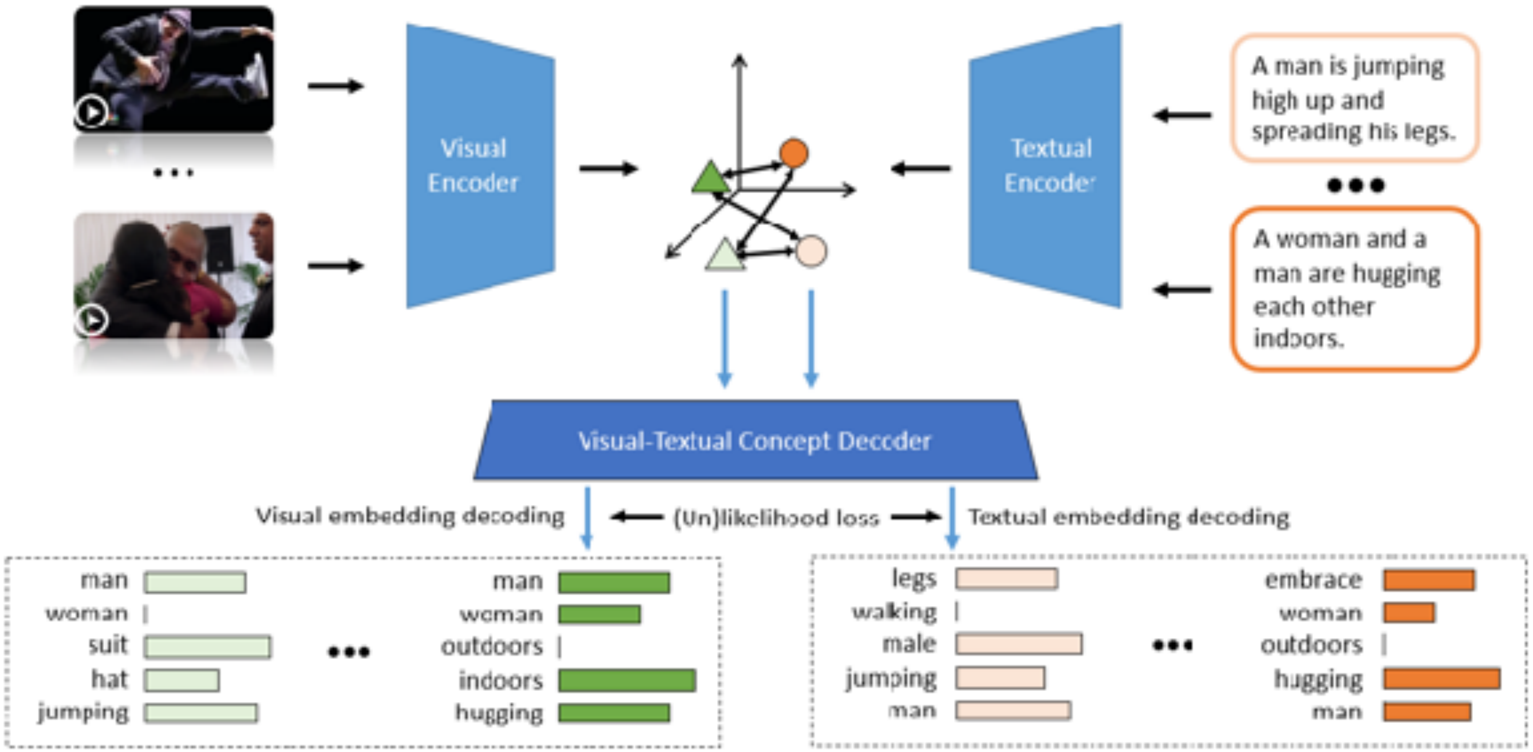
Concept prediction

$\hat{p}$	0.12	0.03	…	0.99	…	0.86	0.03	…	0.96	…	0.21
	truck	cat		man		indoors	outdoors		woman		road

[1] Marneffe et al., Finding Contradictions in Text, ACL, 2008
[2] Welleck et al., Neural text generation with unlikelihood training, ICLR, 2009
[3] Roller et al., Don't say that! making inconsistent dialogue unlikely with unlikelihood training, ACL, 2020

New architecture

- Embedding search, concept search and fusion search



Advantages of the new model

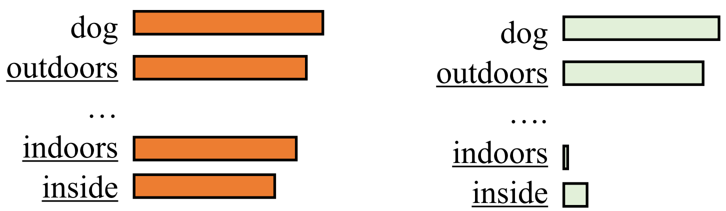
- Make query embedding less sensitive to query formulation
- Likelihood training can address the missing labels problem
- Unlikelihood training avoids frequent and contradicting concepts

a person using sign language

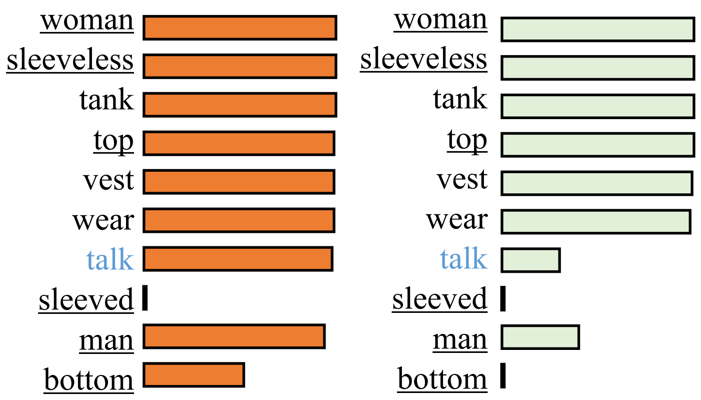
a person communicating using sign language

a <u>woman</u> wearing sleeveless top

Top-15 concepts in the query embedding interpretation: [interpret, communicate, sign, signs, language, deaf, interpreter, asl, use, convey, translate, message, words, show, person]



(a) w/o UL (b) w/ UL



(a) w/o UL (b) w/ UL

Submitted runs on tv21

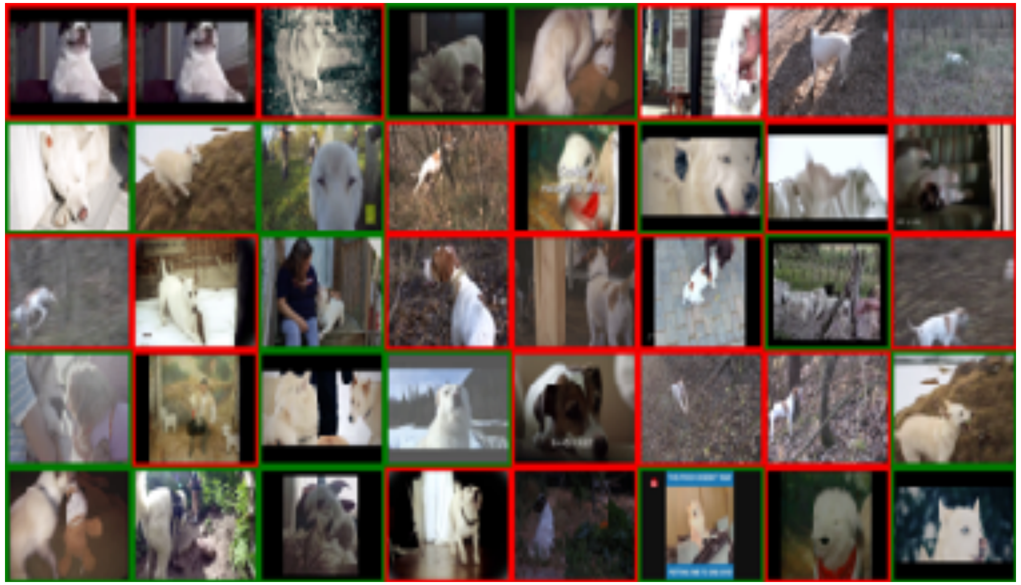
*(+ video features[2,3] + VATEX dataset [4])

Submitted run	Model	Concept search	Embedding search	Fusion search
Baseline #1	Original Dual-task model	0.167	0.167	0.193
Baseline #2	Feature enhancement dual-task model*	0.269	0.278	0.305
Baseline #3	Feature enhancement dual encoding model* [1]	/	0.287	/
RUN1	Phrase model*	0.216	0.301	0.317
RUN2	(Un)likelihood model*	0.270	0.290	0.330
RUN3	RUN1+RUN2	/	/	0.336
RUN4	RUN1+RUN2+Feature enhancement	/	/	**0.355**
Novelty run	Concept searches of RUN1 and RUN2+manual queries	0.297	/	/

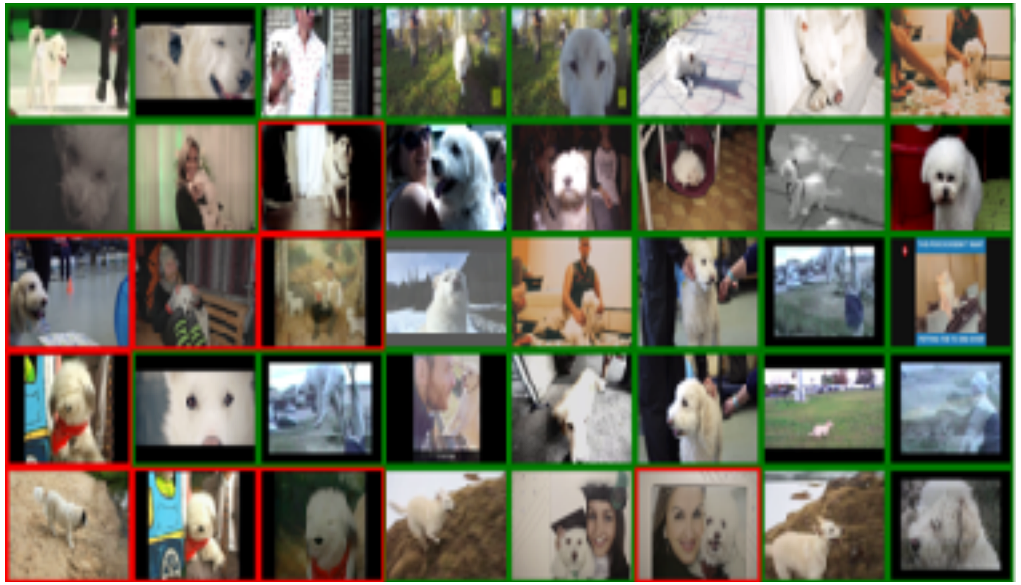
[1] Dong et al., Dual Encoding for Zero-Example Video Retrieval, *CVPR*, 2019
[2] Feichtenhofer et al., Slowfast networks for video recognition, *ICCV*, 2019
[3] Liu et al., Swin transformer: Hierarchical vision transformer using shifted windows, *ICCV*, 2021
[4] Wang et al., Vatex: A large-scale, high-quality multilingual dataset for video-and-language research, ICCV, 2019

Benefit from the phrase vocabulary

- 676 Find shots of a <u>white dog</u>



(a) Dual-task$_{concept}$ (xinfAP=0.167)

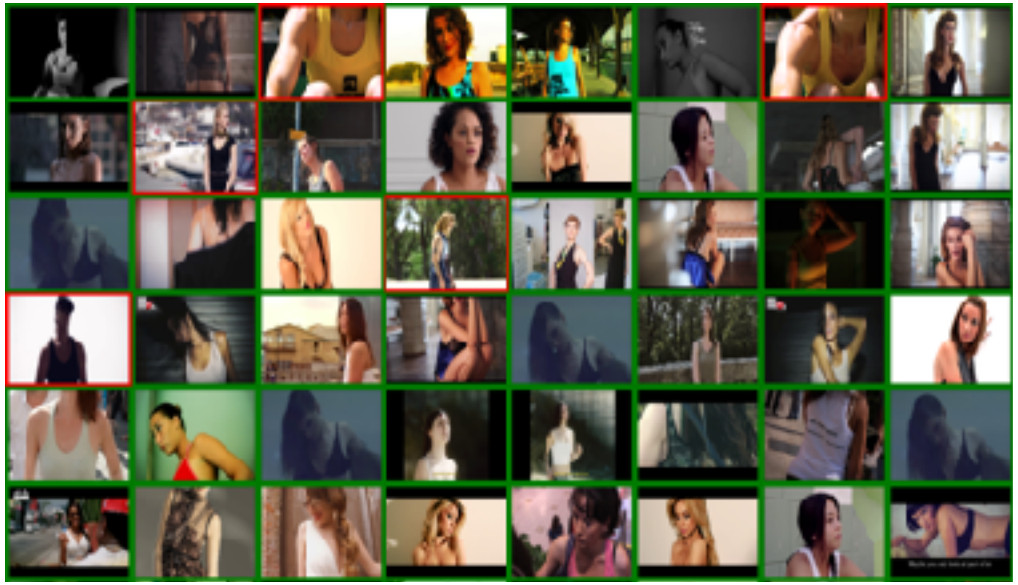


(b) Phrase model$_{concept}$ (xinfAP=0.4252)

Benefit from the unlikelihood training

- 662 Find shots of a <u>woman</u> wearing sleeveless top

(a) Dual-task$_{\text{embedding}}$ (xinfAP=0.355)

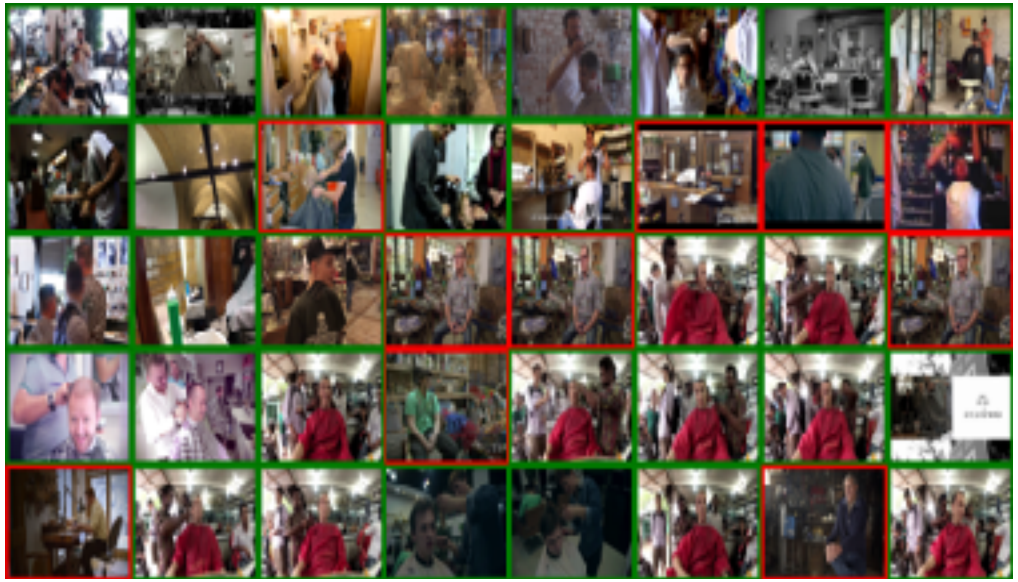


(b) UL model$_{\text{embedding}}$ (xinfAP=0.580)

Suffer from small number of training cases

- 678 Find shots of a man sitting on a barber chair in a shop

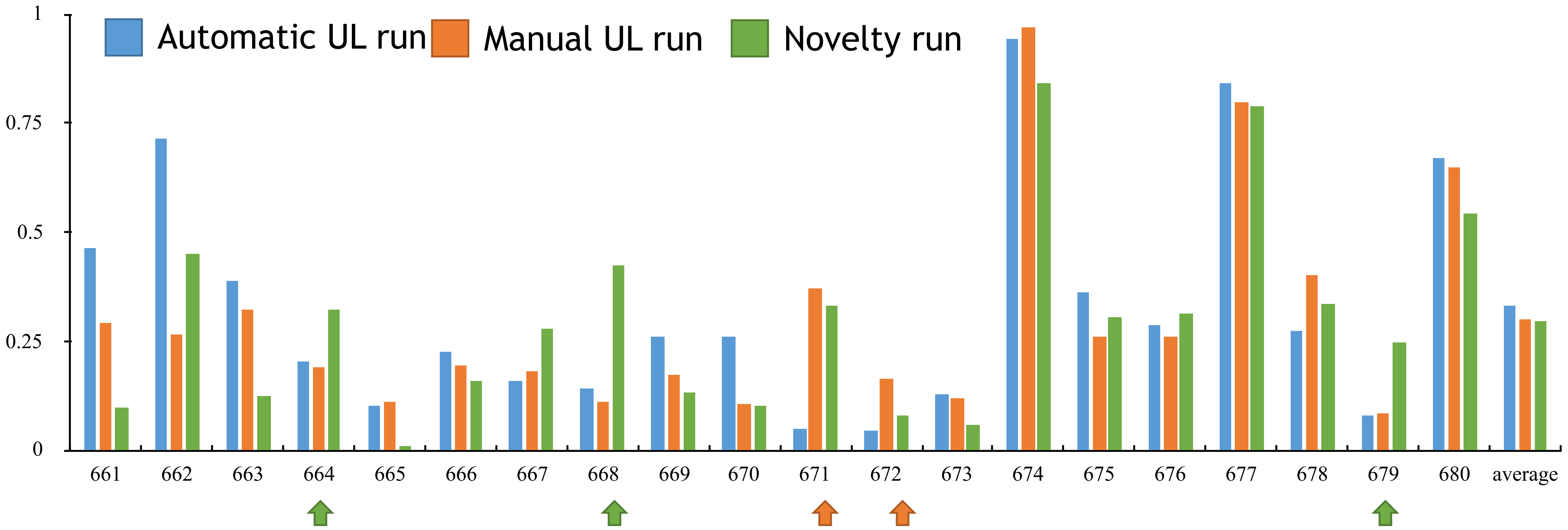
(a) UL model (xinfAP=0.409)



(b) Phrase model$_{concept}$ (xinfAP=0.133)

Automatic Versus Manual (Novelty) runs

- Automatic runs outperform manual runs.
- Manual (Novelty) runs are sensitive to query formulation.



Summary

- Enhanced features and additional dataset significantly improve the performance.
- (Un)likelihood model effectively pull down contradicted videos.
- With phrases, interpretable embeddings are more robust, but concept phrase retrieval rate could be limited by having a small number of training samples.
- Manual runs are sensitive to query formulation and the results are depend on the training data and the video dataset.

Thank you

Q&A